MitoNuc: a database of nuclear genes coding for mitochondrial proteins. Update 2002

Marcella Attimonelli*, Domenico Catalano¹, Carmela Gissi², Giorgio Grillo¹, Flavio Licciulli¹, Sabino Liuni¹, Monica Santamaria, Graziano Pesole² and Cecilia Saccone

Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, via Orabona 4, 70126 Bari, Italy, ¹Centro di Studio sui Mitocondri e Metabolismo Energetico C.N.R., via Orabona 4, 70126 Bari, Italy and ²Dipartimento di Fisiologia e Biochimica Generali, Università di Milano, via Celoria 26, 20122 Milano, Italy

Received September 20, 2001; Accepted September 26, 2001

ABSTRACT

Mitochondria, besides their central role in energy metabolism, have recently been found to be involved in a number of basic processes of cell life and to contribute to the pathogenesis of many degenerative diseases. All functions of mitochondria depend on the interaction of nuclear and organelle genomes. Mitochondrial genomes have been extensively sequenced and analysed and data have been collected in several specialised databases. In order to collect information on nuclear coded mitochondrial proteins we developed MitoNuc, a database containing detailed information on sequenced nuclear genes coding for mitochondrial proteins in Metazoa. The MitoNuc database can be retrieved through SRS and is available via the web site <http://bighost.area.ba.cnr.it/mitochondriome> where other mitochondrial databases developed by our group, the complete list of the sequenced mitochondrial genomes, links to other mitochondrial sites and related information, are available. The MitoAln database, related to MitoNuc in the previous release, reporting the multiple alignments of the relevant homologous protein coding regions, is no longer supported in the present release. In order to keep the links among entries in MitoNuc from homologous proteins, a new field in the database has been defined: the cluster identifier, an alpha numeric code used to identify each cluster of homologous proteins. A comment field derived from the corresponding SWISS-PROT entry has been introduced; this reports clinical data related to dysfunction of the protein. The logic scheme of MitoNuc database has been implemented in the ORACLE DBMS. This will allow the end-users to retrieve data through a friendly interface that will be soon implemented.

INTRODUCTION

Mitochondria, subcellular organelles present in the majority of eukaryotes, contain their own independent genome and

expression machinery. Recently, mitochondria have returned to the interest of researchers since, besides their fundamental role in the cellular energy metabolism, they seem to be involved in a number of central cellular processes such as apoptosis. Moreover, mitochondrial defects contribute to the pathogenesis of many degenerative diseases, to aging and to cancer. Mitochondrial genomes have been extensively sequenced in many organisms and data stored in several specialised databases and web sites [<http://bighost.area.ba.cnr.it/mitochondriome>, GOBASE (1), Entrez Genomes (2)]. A few mitochondrial proteins are coded by the mitochondrial genome. Indeed, most of the mitochondrial proteins are encoded by nuclear genes, synthesised in the cytoplasm and then imported into the organelle. Thus, the cross-talk between the nuclear and the mitochondrial genomes is crucial for mitochondrial biogenesis and function and the two genomes are probably subjected to co-evolutionary processes. The present paper describes the specialised database MitoNuc reporting information on sequenced nuclear genes coding for mitochondrial proteins (*nugemips*) in Metazoa. The complementary database MitoAln, containing multiple alignments of the homologous genes represented in MitoNuc, has no more been supported in the present release. This is because of great difficulty in producing a good multi-alignment when managing distantly related proteins whose function and site features are not always well known. Production of good-quality alignment would require a specific expertise for each protein. In order to keep the links among entries in MitoNuc from homologous proteins, a new field in the database has been defined: the cluster identifier, an alpha numeric code used to identify each cluster of homologous proteins. It is up to the user to produce the multi-alignment once the related homologous sequences have been extracted from the MitoNuc database.

DATABASE STRUCTURE

Each MitoNuc entry defines a nuclear gene coding for a mitochondrial-related protein in a given species. Each entry reports a set of defined information: gene name; gene product name, synonymous and functional classification as defined in the KEYnet database (3); species name and taxonomic classification; Enzyme Classification code (EC) in the case of genes coding for enzymes; metabolic pathways in which the protein product is involved; cellular and sub-mitochondrial

*To whom correspondence should be addressed. Tel: +39 80 5482180; Fax: +39 80 5484467; Email: marcella@area.ba.cnr.it

localisation of the encoded proteins; possible presence of tissue-specific isoforms, cross-references to the EMBL, SWISS-PROT/TrEMBL and UTR databases, comments about clinical data related to protein dysfunction. More than one EMBL nucleotide sequence can be linked to the same MitoNuc entry because of the high level of redundancy of the primary databases where the same gene is frequently reported in different entries. MitoNuc is cross-referenced to each of the redundant entries in order to avoid losing important information related to such redundancy (i.e. polymorphisms, tissue specificity, etc.). The location and description of sequence regions with specific functions present in each of the EMBL entries linked to a given MitoNuc entry has been included in a detailed feature table. In particular, the reported information concerns the protein coding region (CDS), the signal peptide (sig_peptide), the mature peptide (mat_peptide), and the 5' and 3' untranslated regions (5'-UTR and 3'-UTR) of the relevant mRNAs. The MitoNuc Entry does not report any sequence (neither the nucleotide nor the protein sequence); however, these data can be extracted thanks to the linking database function implemented in SRS.

The MitoNuc database is formatted in an EMBL-like flat-file format, thus allowing the indexing of the database in SRS (4). The logic schema of the database has been produced (<http://bighost.area.ba.cnr.it/BIG/Tutorials/Mitonuc/Dbschema.html>) and implemented in ORACLE on the basis of the above described data fields. This is the first stage of the improvements we have planned to apply to the MitoNuc database in order to facilitate the end-users in data retrieval and analysis. With this goal we are also planning to interface as soon as possible the MitoNuc database to Gene Ontology (GO) (5).

DATABASE DATA SOURCE

Data sources for MitoNuc are SWISSALL (6) and EMBL (7) databases. Data are retrieved from SWISS-PROT through SRS searching for METAZOA and MITOCHONDRION. Data selected are thoroughly analysed and classified in gene clusters on the basis of their function as annotated in SWISS-PROT. Additional information concerning the gene and the gene product have been obtained from the literature or extracted from specialised protein databases. The MitoNuc feature table is produced through an automatic browsing of SWISS-PROT, EMBL and UTRdb (8) database feature tables, followed by a human revision of the obtained results.

MitoNuc DATA CONTENT

MitoNuc database is updated to the SWISS-PROT release 40 (February 2001) reporting 951 entries with cross-references to a total of 1820 EMBL entries. Such entries refer to 53 different metazoan species, most of them, 732 entries, belonging to the mammalian class and 262 entries related to human sequences; 897 entries of MITONUC entries are grouped in clusters for a total of 239 clusters. The remaining 54 entries report proteins whose sequence has been produced in one species only.

AVAILABILITY AND RETRIEVAL OF THE DATABASES

MitoNuc is available on the web through the 'mitochondrome' web site and can be queried using the SRS retrieval system

(<http://bighost.area.ba.cnr.it/srs>). Specific sequence regions of the MitoNuc entry defined in the feature table can be easily selected and extracted, thus greatly helping further specific sequence analyses.

CONCLUSIONS AND PERSPECTIVES

MitoNuc database can greatly help in a better knowledge of structure, evolutionary and functional features of *nugemips* and their products, thus providing an important contribution to devise suitable therapies against diseases associated with mitochondria, which have grown tremendously lately. Moreover, the availability of the complete human genome can greatly contribute to locate MitoNuc *nugemips* on the genome in order to contribute to disease association studies.

The MITODAT (9) (<http://www-lecb.ncifcrf.gov/mitoDat>) and MITOP (10; <http://www.mips.biochem.mpg.de/proj/medgen/mitop/>) data collections contain similar information on mitochondrion-related proteins, but they are restricted to a more limited species sample (only human for MITODAT and only five species for MITOP), and do not allow sequence extractions and easy information retrieval. Our databases have been structured in order to allow specific selections, combining various criteria and easy extraction of sequence data, also limited to regions of specific interest.

ACKNOWLEDGEMENTS

This work has been supported by 'Ministero Università e Ricerca Scientifica', Italy (PRIN99, Programma Biotecnologie legge 95/95-MURST 5%; Progetto MURST Cluster C03/2000, CEGBA).

REFERENCES

- Shimko, N., Liu, L., Lang, B.F. and Burger, G. (2001) GOBASE: the organelle genome database. *Nucleic Acids Res.*, **29**, 128–132.
- Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L. and Rapp, B.A. (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **29**, 11–16. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 13–16.
- Catalano, D., D'Elia, D., Licciulli, F. and Attimonelli, M. (2000) Update of KEYnet: a gene and protein names database for biosequences functional organisation. *Nucleic Acids Res.*, **28**, 372–373.
- Etzold, T., Ulyanov, A. and Argos, P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
- Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
- Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Lombard, V., Lopez, R., Parkinson, H. *et al.* (2001) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **29**, 17–21. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 21–26.
- Pesole, G., Liuni, S., Grillo, G., Licciulli, F., Larizza, A., Makalowski, W. and Saccone, C. (2000) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.*, **28**, 193–196. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 335–340.
- Lemkin, P.F., Chipperfield, M., Merrill, C. and Zullo, S. (1996) A World Wide Web (WWW) server database engine for an organelle database, MitoDat. *Electrophoresis*, **17**, 566–572.
- Scharfe, C., Zaccaria, P., Hoernagel, K., Jaksch, M., Klopstock, T., Lill, R., Prokisch, H., Gerbitz, K.D., Mewes, H.W. and Meitinger, T. (2000) MITOP, the mitochondrial proteome database: 2000 update. *Nucleic Acids Res.*, **28**, 155–158.